Eur J Med Chem (1996) 31, 479–484

Quantitative electronic structure–activity relationships of pyridinium cephalosporins using nonparametric regression methods

V Nguyen-Cong, BM Rode

Theoretical Chemistry Division, Institute for General, Inorganic and Theoretical Chemistry, University of Innsbruck, 52a Innrain, 6020 Innsbruck, Austria

(Received 26 October 1995; accepted 26 January 1996)

Summary — Projection pursuit regression (PPR) was applied to interpret and predict the antibacterial activity of pyridinium cephalosporins using semiempirical quantum mechanical descriptors. This method can deal with responses due to interactions of predictors (descriptors) which cannot be completely represented by additive regression models. Based on leave-one-out cross-validation, the best PPR model gave a cross-validated r^2 or q^2 value of 0.711, whereas the traditional method, multiple linear regression, and another additive nonparametric model, alternating conditional expectations, produced the best q^2 values: 0.233 and 0.324 respectively. Its ability to provide models with good predictive ability reveals that PPR is a valuable tool in quantitative structure–activity relationship studies.

quantitative structure–activity relationship / quantitative electronic structure–activity relationship / multiple linear regression / alternating conditional expectations / projection pursuit regression

Introduction

Quantitative structure–activity relationships (QSAR) are based on the assumption that the biological activities of a chemical compound are related to, and hence characterizable by, some of its physicochemical parameters such as solubility, lipophilicity, polarity and steric structure [1]. A method that has led to many successful QSAR studies is Hansch analysis [2–4]. Quantitative electronic structure–activity relationship (QESAR) analysis is an alternative to the QSAR concept, assuming that the biological activities of a chemical compound can be described by its electronic molecular parameters. Recent studies ([5], Texler, Nguyen-Cong and Rode, unpublished results) have shown some success in applying multiple linear regression to building relationships between atomic net charges and biological activities of natural cytotoxic and antibiotic drugs.

If linearity in the dependence of the responses on the predictor variables prevails, linear regression is extremely convenient and useful because it supplies a simple description of the data, weighs the contribution of each predictor with a single coefficient, and provides a simple method for predicting new observations. However, this assumption of linearity does not always hold in QSAR studies, and thus some nonlinear methods [6, 7] were proposed as their performance was expected to be better than linear regression in such cases.

In a continuation of recent studies [8, 9] focusing on QESAR using nonparametric nonlinear regression methods, this work applies the projection pursuit regression (PPR) method to studying antibiotic drugs of cephalosporin type. PPR was originally proposed by Friedman and Stuetzle [10] for a single response variable. Later on, Friedman [11] presented a more efficient algorithm, suitable for multiple response regression and classification. The PPR model is also a special case of a feed-forward neural network with one hidden layer for supervised learning [12, 13].

Materials and methods

Data

A series of 27 pyridinium cephalosporins related to cefpirome (fig 1) was used in this work. These compounds were synthesized and tested by Latrell et al [14]. The in vitro activity against *Staphylococcus aureus* determined by a serial dilution test was expressed as minimum inhibitory concentration (MIC). The chemical structures and corresponding

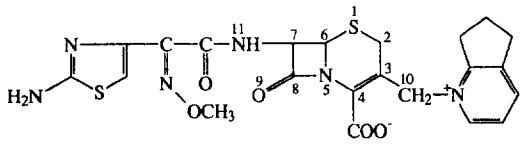


Fig 1. Structure of cefpirome.

antibiotic activities are shown in table I. The molecular geometries of the studied compounds were optimized using the MM+ force field, and the atomic net charges (table II) in the essential part of this class of compounds, corresponding to the predictors in this study, were evaluated for their optimized structures by the semiempirical molecular orbital method PM3 and Mulliken population analysis. All calculations were carried out with HyperChem software (HyperChem, Hypercute Inc, Ontario, Canada).

Computational procedure

The problem common to QSAR studies is that of adequately approximating a function of a number of variables, based only on observational data. Given n observations of a response variable Y and a set of predictor variables $X_j (j = 1, 2, \ldots, p)$,

$$\{(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})\}_{i=1}^{n}, \quad (1)$$

the goal is to model the dependence of Y on $X_1, X_2, \ldots, X_p$. The relationship is presumed to be described by

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon \quad (2)$$

where f is a single-valued function of p variables and ε is a random variable with zero expectation, $E[\varepsilon] = 0$.

Alternating conditional expectations

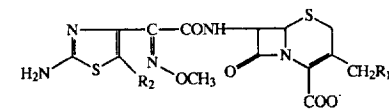
Alternating conditional expectations (ACE) is an additive regression model introduced by Breiman and Friedman [15]. It has the form

$$\theta(Y) = \phi_1(X_1) + \ldots + \phi_p(X_p) + \varepsilon = \sum_{j=1}^{p} \phi_j(X_j) + \varepsilon \quad (3)$$

where Y and $X_j (j = 1, 2, \ldots, p)$ have a joint distribution, θ and ϕ_j are nonlinear transformations with the constraints $E\theta = 0$ and $E\phi_j = 0$. ACE applies an iterative procedure for finding optimal transformations that minimize the fraction of variance unexplained

$$e^2 = \frac{\sum_{i=1}^{n} \left[\hat{\theta}(y_i) - \sum_{j=1}^{p} \hat{\phi}_j(x_{ij})\right]}{\sum_{n} \hat{\theta}^2(y_i)} \quad (4)$$

Table I. Chemical structure and antibiotic activity of pyrimidinium cephalosporins[a].



No.	R_1	MIC	No.	R_1	MIC
1	Cefpirome sulfate	0.31	15	$C(CH_3)_3$	3.13
2	(sulfate)	0.39	16	(sulfate)	0.31
3		0.78	17		0.78
4	(sulfate)	0.31	18		0.78
5		0.62	19	OCH_3 (sulfate)	0.31
6		0.31	20	OC_2H_5	0.39
7		0.62	21	OC_3H_7	0.39
8		0.78	22	SCH_3	0.39
9		0.62	23	F	0.39
10		0.19	24	Cl	0.78
11		1.56	25	$CONH_2$	0.39
12	CH_3	1.56	26	$COCH_3$	0.78
13	C_2H_5	0.78	27	SO_2NH_2	3.13
14	C_3H_7	1.56			

[a]Compound **9** R_2 = Cl, compound **10** R_2 = Br, and the remaining compounds R_2 = H.

where $\hat{\theta}$ and $\hat{\phi}_j$ are estimates of θ and ϕ_j. In other words, ACE provides nonlinear transformations of both predictor and response variables to maximize the correlation between the transformed responses and the sum of the transformed predictors. The ACE transfor-

Table II. Atomic net charges used in this study (see fig 1).

	S1	*C2*	*C3*	*C4*	*N5*	*C6*	*C7*	*C8*	*O9*	*C10*	*N11*
1	0.022363	–0.112224	–0.370443	0.101838	0.091978	–0.147615	–0.075786	0.246555	–0.244126	–0.138300	–0.021840
2	0.018459	–0.111602	–0.363304	0.098542	0.092998	–0.151726	–0.075267	0.245882	–0.244999	–0.143136	–0.021693
3	0.024448	–0.111717	–0.374636	0.104218	0.092280	–0.145598	–0.076243	0.246314	–0.243317	–0.126829	–0.021196
4	0.022995	–0.111287	–0.371881	0.102461	0.092704	–0.147552	–0.075993	0.246223	–0.243304	–0.128305	–0.022013
5	0.022837	–0.110921	–0.374426	0.103782	0.093013	–0.148780	–0.076332	0.245777	–0.242804	–0.120065	–0.022037
6	0.022868	–0.111947	–0.370992	0.103042	0.092454	–0.145860	–0.076600	0.245970	–0.244011	–0.130034	–0.021118
7	0.023959	–0.115042	–0.369864	0.101699	0.090810	–0.143483	–0.076585	0.247533	–0.244041	–0.131484	–0.021598
8	0.026121	–0.114050	–0.375070	0.104499	0.092642	–0.144819	–0.077621	0.245850	–0.241461	–0.123380	–0.020638
9	0.017917	–0.110205	–0.371871	0.102636	0.093004	–0.152684	–0.074359	0.246362	–0.242125	–0.136628	–0.009665
10	0.024670	–0.113717	–0.373625	0.103692	0.090846	–0.143659	–0.074734	0.247804	–0.239519	–0.138456	–0.004342
11	0.027414	–0.115398	–0.378103	0.108522	0.091284	–0.143679	–0.076318	0.248101	–0.241952	–0.125958	–0.021649
12	0.023872	–0.113671	–0.374804	0.106509	0.091496	–0.145706	–0.075841	0.248348	–0.243974	–0.124966	–0.022088
13	0.027284	–0.115876	–0.376933	0.106831	0.092492	–0.144569	–0.076499	0.246373	–0.241745	–0.123947	–0.021784
14	0.025621	–0.114205	–0.376011	0.106322	0.092731	–0.146202	–0.076231	0.246315	–0.242249	–0.124989	–0.021967
15	0.019408	–0.112731	–0.371772	0.105203	0.093186	–0.151881	–0.074889	0.247318	–0.245108	–0.126501	–0.023293
16	0.027752	–0.116535	–0.376169	0.106274	0.091934	–0.142634	–0.077798	0.245825	–0.240611	–0.124853	–0.019618
17	0.021141	–0.112576	–0.371758	0.104939	0.093133	–0.150072	–0.075202	0.246548	–0.244287	–0.127374	–0.023151
18	0.023190	–0.112999	–0.373114	0.103947	0.091556	–0.146114	–0.076940	0.247352	–0.243835	–0.123171	–0.021529
19	0.024945	–0.112382	–0.372939	0.103757	0.093949	–0.149027	–0.075988	0.244932	–0.242254	–0.133214	–0.022167
20	0.023482	–0.114445	–0.375371	0.105675	0.090574	–0.146545	–0.076524	0.249046	–0.244144	–0.116586	–0.022000
21	0.022717	–0.112844	–0.376126	0.106091	0.090569	–0.147078	–0.076277	0.249000	–0.244718	–0.116020	–0.022010
22	0.025561	–0.112358	–0.372388	0.103501	0.092821	–0.147877	–0.075741	0.245429	–0.242584	–0.132065	–0.021561
23	0.027568	–0.113945	–0.376642	0.110155	0.089884	–0.144655	–0.074786	0.249766	–0.243789	–0.131275	–0.022710
24	0.025253	–0.113912	–0.375504	0.109067	0.090076	–0.144852	–0.075428	0.249490	–0.244454	–0.128393	–0.022420
25	0.031814	–0.115407	–0.379188	0.107854	0.092233	–0.145700	–0.075686	0.245528	–0.238757	–0.128070	–0.020962
26	0.030299	–0.112368	–0.378614	0.106957	0.092507	–0.149516	–0.073993	0.244919	–0.239424	–0.127790	–0.022073
27	0.030713	–0.114121	–0.381087	0.110857	0.090735	–0.145544	–0.074841	0.246653	–0.240223	–0.132011	–0.021820

mations are obtained using a two-dimensional scatterplot smoother called Supersmoother [16, 17]. Supersmoother is a variable-span smoother based on local linear fits, with the optimal span chosen by local cross-validation [18].

Although ACE is one of the most powerful tools for data analysis, providing knowledge of transformations needed to interpret and understand the relationship between the response and each of the predictors, it may cause some anomalous results, and thus lead to inappropriate interpretations [19, 20]. Some of those anomalous effects are: i) when there is little or no relationship between predictor variables and the response, ACE may result in strong-looking transformations, mostly in the quadratic form and maximal correlations up to about 0.3; ii) when there are disjoint clusters of data, ACE may result in step functions as optimal transformations; iii) ACE transformations can change abruptly even as the underlying distribution of the data changes very slightly; and iv) there might exist many solutions that give nearly the same correlations but which are nonunique transformations.

Projection pursuit regression

Projection pursuit regression (PPR) [11, 12] models the response variable by a linear combination of predictor functions f_m, where the f_m are estimated using the supersmoother.

$$\hat{Y} = \bar{Y} + \sum_{m=1}^{M} \beta_m f_m(\alpha_m^T X) \qquad (5)$$

with $\bar{Y} = EY = \frac{1}{n}\sum_{i=1}^{n} y_i$, $Ef_m = 0$, $Ef^2_m = 1$ and $\sum_{j=1}^{p} \alpha_{mj}^2 = 1$.

The term projection pursuit originated from the use of a numerical optimization technique to find direction vectors $\alpha_m^T = (\alpha_{m1}, \ldots, \alpha_{mp})$ so that $\alpha^T X$ represents the projection of X in the direction α. The model parameters, the projection directions α_m^T, the response linear combinations β_m^T and the functions f_m in equation (7) are estimated to minimize the mean squared error.

In contrast to additive models like ACE, PPR can model the effect of interactions between the predictor

variables. This is extremely useful in QESAR studies because atomic net charges, especially of neighbouring centres in a molecule, are necessarily interdependent. A simple example commonly used to demonstrate PPR's ability to model interaction effects is to consider the function $Y = X_1X_2$. It is clear that additive models cannot represent this multiplicative relationship. However, Y can be described by a PPR model with $\bar{Y} = 0$, $M = 2$, $\beta_1 = \beta_2 = 1/4$, $\alpha_1^T = (1, 1)$, $\alpha_2^T = (1, -1)$, $f_1(x) = x^2$ and $f_2(x) = -x^2$.

The algorithm for a single response variable is summarized as follows:

1. Initialize term index ($m = 1$) and response variable residuals $\left(e_{0,i} = y_i \frac{1}{n} \sum y_i\right)$.

2. Use a trial direction vector α to construct a linear combination $\alpha^T X$ and fit a smooth curve f to estimate e_{m-1} from $\alpha^T X$ using the supersmoother.

3. With f fixed, find the vector α_m that minimizes the sum of squares

$$\sum_{i=1}^{n} \left(e_{m-1,i} - f\left(\alpha_m^T X_i\right)\right)^2$$

and fit a smooth curve f_m to estimate e_{m-1} from $\alpha_m^T X$. Let $e_{m,i} = e_{m-1,i} - f_m(\alpha_m^T X_i)$.

4. Increase the term index ($m = m + 1$) and go to Step 2 until the residual vector e_m is smaller than a user-specified threshold or m has reached a present value M.

In order to avoid local minima, more terms than necessary ($M_{max} > M$) can be specified and a backward stepwise model selection applied to prune the model back to a total of M terms.

Results and discussion

Multiple linear regression

Application of multiple linear regression (MLR) to all possible subsets of eleven predictors as characterized in figure 1 and natural logarithms of MIC values, resulted in a model containing seven predictors (S1, C3, C6, C7, O9, C10, N11) with a maximum value of 0.233 for the q^2 value [21]; the corresponding multiple r^2 of this model was 0.590. These values reveal that the fitting and predictive ability of a linear model is generally poor. Thus, a nonlinear relationship between the activity and molecular electronic parameters had to be expected for this data set.

Alternating conditional expectations

Using the same subset selection procedure for the best linear model, the ACE technique gave a six-predictor model (C2, C3, C7, C8, O9, C10) with 0.324 and 0.986 for q^2 and r^2 ($r^2 = 1 - e^2$) respectively. It can be seen that ACE not only improves both fitting and prediction, but also uses a lower number of predictors. However, it is noted that the predictive ability expressed by the q^2 value is quite low compared to its quality of fitting, and thus this model is probably a chance correlation.

Projection pursuit regression

Table III gives the optimal PPR model for the dataset of 27 cephalosporins. The table includes the number of terms in the model (M), the number of predictors (p), the multiple r^2 value, and the predictive ability (q^2). The model includes six predictors (S1, C2, C3, C8, C10 and N11), accounting for 92.1% of its variance. It predicts significantly better than the ACE model due to a high value of q^2. This model is generally in agreement with previous statements about structure-related biological activity of cephalosporins. It has been reported that the nature and spatial orientation of the 7β-acylamino side chain strongly influences the β-lactamase stability and binding to penicillin-binding proteins. Furthermore, the substituents at the C3 and C10 positions affect the antibacterial activity, metabolic stability, pharmacokinetics and adverse effects [22]. Figure 2 represents the scatterplot of fitted and predicted versus observed activity values. The corresponding residuals in figure 3 do not exhibit any unusual structure. The two estimated predictor functions f_m, corresponding to the two terms, are shown in figures 4a and 4b by plotting the linear combination value versus the function value for 27 compounds. Both plots are substantially nonlinear, suggesting that the antibacterial activity of pyridinium cephalosporins depends on their atomic net charges in a complex manner.

It is also useful to have an idea of how strongly each predictor enters into the final model. The relative importance measure of the predictor X_j to a PPR model [11] is defined as

$$I_j = \sigma_j E \left| \sum_{m=1}^{M} \beta_m \alpha_{jm} f'(\alpha_m^T X) \right| \qquad (10)$$

Table III. The best PPR model for the cephalosporin dataset.

M	p	r^2	q^2
2	6	0.921	0.711

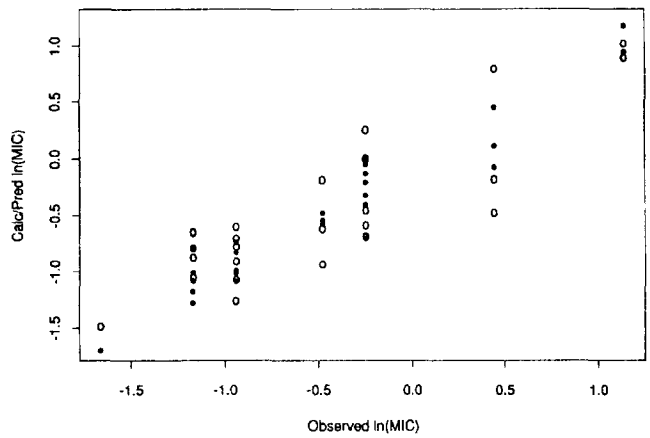


Fig 2. Observed ln (MIC) values plotted against calculated (solid circles) and predicted (open circles) ln (MIC) values.

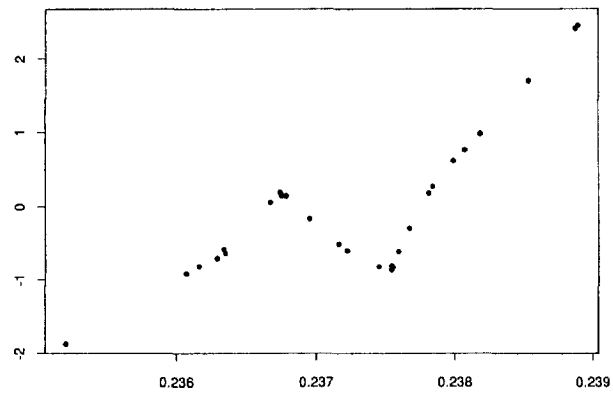


Fig 4a. Term 1 predictor function.

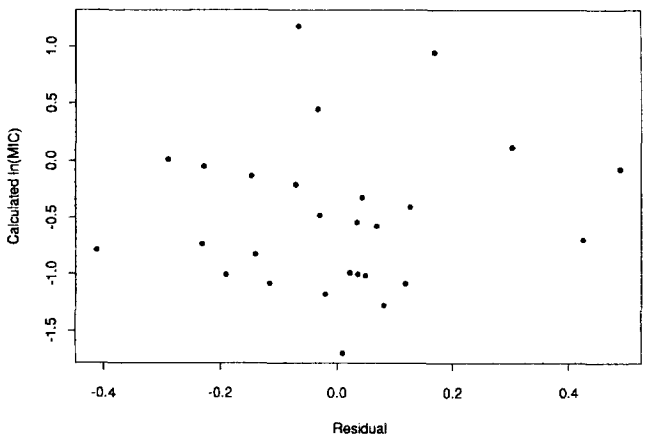


Fig 3. Residuals plotted against calculated ln (MIC) values.

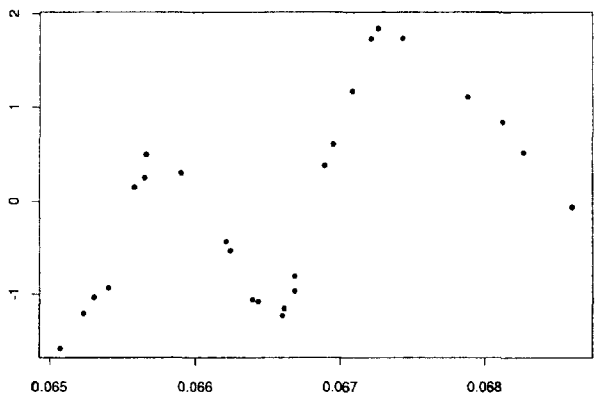


Fig 4b. Term 2 predictor function.

where σ_j is a scale measure of the predictor X_j, and $f'_m(z) = df_m/dz$. For $M = 1$, I_j becomes $\sigma_j|\alpha_j|$, similar to the measure of linear models. Table IV presents the relative importance of each predictor (atomic net charge) within the final model. These data clearly confirm the sensitivity to C3 and C10 substitutions observed by Dürckheimer et al [22] and further indicate that S1 plays the most important role among the heteroatoms in the cephalosporins, followed by N11.

Comparison of the predictor variables of the three methods applied shows that MLR and PPR give quite similar indications of the importance of heteroatoms in the active centre of the drug, whereas ACE focuses on the carbon atoms of the backbone. This indicates that MLR and PPR are more suitable for predicting the role of central atoms in the interaction of the drug with its receptor. Since MLR is apparently not satisfactory for the particular case of cephalosporin drugs, this interaction is most probably of a more complicated nature than in the case of other, simpler drugs. Any attempt to design new drugs of cephalosporin type will have to take into account this finding, and could be based on the result of the PPR study. The relative importance of atoms resulting from the PPR study indicates that the interaction consists of both hydrophobic and hydrogen-bonding interactions of the drug's central ring system. The receptor structure is expected, therefore, to be rather complex and consisting of several different functional groups.

Table IV. Contribution of predictor variables to the final PPR model.

Predictor	*Importance*
C3	1.00
S1	0.96
C10	0.41
N11	0.35
C2	0.29
C8	0.17

In conclusion, it can be stated that projection pursuit regression represents a superior approach for cases where nonlinear QSARs exist and the form of the relationship is unknown. Since this can be expected for the majority of QSARs, PPR may become a fairly common tool for such investigations in the future.

Acknowledgments

A grant from the Austrian Federal Government for Vu Nguyen-Cong is gratefully acknowledged. One of the authors (V Nguyen-Cong) would like to thank Ho-Thi Cam-Hoai for personal assistance. Support from JH Friedman, Department of Statistics Stanford University, who made the Fortran subroutines of ACE and PPR available, and from GUH Seeber, Institute for Statistics University of Innsbruck, through helpful discussions, is also acknowledged.

References

1 Silverman RB (1992) *The Organic Chemistry of Drug Design and Drug Action*. Academic, San Diego
2 Hansch C, Fujita T (1963) *J Am Chem Soc* 86, 1616–1626
3 Kubinyi H (1993) QSAR: *Hansch Analysis and Related Approaches*. VCH, Weinheim
4 Hansch C, Leo A (1995) *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*. American Chemical Society, Washington DC
5 Dang GV, Rode BM, Stuppner H (1994) *Eur J Pharm Sci* 2, 331–350
6 Clare BW (1995) Alternating conditional expectations in QSAR. In: *Advanced Computer-Assisted Techniques in Drug Discovery* (van de Waterbeemd H, ed) VCH, Weinheim, 281–292
7 Zupan J, Gasteiger J (1993) *Neural Networks for Chemists: An Introduction*. VCH, Weinheim
8 Nguyen-Cong V, Rode BM (1995) *J Chem Inf Comput Sci*, in press
9 Nguyen-Cong V, Rode BM (1995) *Quant Struct-Act Relat*, in press
10 Friedman JH, Stuetzle W (1981) *J Am Stat Assoc* 76, 817–823
11 Friedman JH (1985) *Classification and multiple regression through projection pursuit. Technical Report No 12*. Department of Statistics, Stanford University
12 Hertz J, Krogh A, Palmer RG (1992) *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, CA
13 Hwang JN, Lay SR, Maechler M, Martin D, Schimert J (1994) *IEEE T Neural Networks* 5, 342–353
14 Latrell R, Blumbach J, Duerckheimer W et al (1988) *J Antibiotics* 41, 1374–1394
15 Breiman L, Friedman JH (1985) *J Am Stat Assoc* 80, 580–619
16 Friedman JH, Stuetzle W (1982) *Smoothing of scatterplots. Technical Report ORION006*, Department of Statistics, Stanford University
17 Friedman JH (1984) *A variable span smoother. Technical Report LCS5*, Department of Statistics, Stanford University
18 Stone M (1974) *J R Stat Soc Ser B* 36, 111–147
19 Buja A (1990) *Ann Stat* 18, 1032–1069
20 Hastie TJ, Tibshirani RJ (1990) *Generalized Additive Models*. Chapman and Hall, London
21 Cramer III RD , DePriest SA, Patterson DE, Hecht P (1993) The developing practice of comparative molecular field analysis. In: *3D QSAR in Drug Design. Theory, Methods and Applications* (Kubinyi H, ed) ESCOM, Leiden, 443–485
22 Dürckheimer W, Adam F, Fischer G, Kirrstetter R (1988) Recent developments in the field of cephem antibiotics. In: *Advances in Drug Research Vol. 17* (Testa B, ed) Academic, London, 61–234